Predicting Student Academic Success and Engagement Using Bayesian Networks

Iqra Ahmed, Mahrukh Yousuf, Sabahat Zahra

***Abstract*—This study investigated the application of Bayesian Networks to predict student academic success and engagement graduate-level students. By integrating data-driven and expert-defined factors, including prior academic achievements, personal habits, social support, mental health and institutional quality, the model captures the interdependencies influencing academic performance. The network is trained and validated on real-world data using Maximum Likelihood Estimation for parameter learning. Variable Elimination is employed for inference. This work aims to highlight the potential of Bayesian Networks in identifying key determinants of student success to offer insights for educators and administrators to foster academic excellence.**

I. Introduction

Student academic success are fundamental indicators of educational quality and institutional impact. These outcomes depend on a complex interaction of individual, familial, and institutional factors, such as prior academic performance, social support and personal habits. Traditional methods of analysis often struggle to capture the intricate interdependencies among these variables. Bayesian Networks provide a probabilistic framework to model such relationships, allowing for effective prediction and inference. In this study, we implement a Bayesian Network to predict student academic success and engagement at universities, uniquely suited to address this challenge. The challenge laid in modelling the complex interdependencies among these factors. They allow for representation of causal relationships, handling uncertainty and providing interpretable predictions. This report explores the development of out Bayesian Network to predict academic success. focusing on key determinants such as attendance, motivation levels, tutoring sessions, and access to resources. By using real-world data, we aim to identify critical factors influencing student outcomes and provide stakeholders with practical recommendations for targeted interventions.

II. Literataure Review

The use of sophisticated statistical and machine learning models has been motivated by the growing complexity of educational systems and the complicated nature of student performance. In this area, Bayesian Networks (BNs) are very useful since they make it possible to depict probabilistic correlations between variables and produce results that are easy to understand. The methods, conclusions, and ramifications for higher education of the major research that have used BNs to model academic achievement and engagement are summarized in this study.

Bayesian Networks in Education
Bayesian Networks have been widely recognized for their ability to capture the causal relationships between diverse factors influencing academic performance. Previous studies have demonstrated the effectiveness of probabilistic graphical models in educational research. For instance, Pearl,2000 [1] pioneered the use of Bayesian Networks as a method for causal reasoning and probabilistic inference.

Previous studies have demonstrated the effectiveness of probabilistic graphical models in educational research. D. Karaboğa and E. Demir [2] employed PISA data to model academic success, identifying critical determinants such as parental support, resilience, and teacher quality. Their research underscored the capacity of Bayesian Networks to model both direct and indirect effects, such as how family income impacts access to educational resources, which in turn influences student performance.

Factors Influencing Student Performance Probabilistic graphical models, particularly Bayesian Networks (BNs), have emerged as a powerful tool for modeling complex educational systems. Chen and Cheng [2] demonstrated the effectiveness of BNs in capturing interdependencies between learning-related variables, highlighting their superiority over traditional statistical methods in handling uncertainty and representing causal relationships. Multiple critical factors identified are:

1) Individual Factors: Research by Pintrich and Schunk [3] and Ryan and Deci [4] highlights the critical role of individual characteristics. Motivation levels are widely recognized as critical determinants of academic performance, influencing a student's ability to engage and excel in their studies. Learning disabilities and individual learning styles also significantly shape educational outcomes, as they necessitate tailored instructional approaches to support diverse needs, according to Kolb [5] and Dunn and Dunn [6]. Furthermore, study hours and sleep patterns exhibit intricate interrelations with academic achievement, where an optimal balance between effort and rest is essential for cognitive function and sustained performance. These individual factors collectively underscore the multifaceted nature of academic success.

2) Socio-Economic Factors: Extensive research by

Coleman et al. [7] and subsequent studies illuminate the impact of socio-economic variables. Fan and Chen's meta-analysis [8] demonstrates a strong correlation between parental engagement and student achievement, this highlights importance of parents' roles in fostering a supportive environment that promotes learning, motivation, and better performance in academic settings. Such findings reinforce the need for collaborative efforts between educators and families to enhance student success. Reardon [9] and Duncan and Magnuson [10] highlight how socio-economic status influences educational opportunities and outcomes, demonstrating how disparities in resources and support systems significantly impact academic achievement.

3) Institutional Factors: Institutional contexts significantly shape student performance through multiple dimensions. Teacher quality has a profound impact on learning outcomes, as highlighted by Duncan and Magnuson [10] and Fan and Chen's [8], who emphasize the role of effective educators in enhancing academic success. School type and resources also play a pivotal role, with Lubienski and Lubienski [11] illustrating how institutional support and resource availability directly affect student outcomes. Additionally, Reardon [9] demonstrates the strong influence of peer relationships, showing how social connections foster motivation and positively influence achievement. Together, these factors highlight the interplay of institutional support in education

Causal Inference and Variable Interaction

Nunes and Lemaire [12] highlighted the distinct advantages of Bayesian Networks in educational research. They demonstrated how BNs identify both direct and indirect factors influencing academic performance, offering a structured view of their interdependencies. Furthermore, BNs quantify probabilistic relationships between variables, enabling precise modeling of their impact. Compared to traditional statistical approaches, BNs provide more nuanced and interpretable insights, making them particularly suited for analyzing complex systems like education. This reinforces their utility in exploring multifactorial influences on student success.

Applications in Higher Education:

In the context of higher education, Karaboğa and Demir [2] highlighted the transformative role of Bayesian Networks in education, showcasing their ability to generate actionable insights for targeted interventions. By modeling individualized factors, BNs support personalized learning strategies tailored to students' unique needs. Additionally, the study emphasized the utility of BNs in capturing the complexities of educational ecosystems, offering a comprehensive understanding of interrelated variables. These contributions underscore the growing importance of probabilistic methods in advancing evidence-based educational practices.

Research Limitations and Future Directions:

Despite their advantages, Bayesian Networks face challenges, including computational complexity, sensitivity to data quality, and the need for extensive domain expertise to define causal relationships. To overcome these limitations, recent research advocates for hybrid approaches that combine Bayesian inference with machine learning techniques like feature selection algorithms and ensemble methods

Implications for This Study:

The reviewed literature highlights the effectiveness of Bayesian Networks in understanding and predicting educational outcomes. Drawing on these findings, this study aims to develop a Bayesian Network integrating variables such as attendance, motivation, access to resources, and mental health. By leveraging a combination of data-driven and expert-defined relationships, this model seeks to provide actionable insights for fostering academic success and engagement.

III. Methodology

A. Model Development

The development of the Bayesian Network followed a hybrid approach, integrating domain expertise with data-driven techniques to capture the dependencies and probabilistic relationships among variables effectively.

1) Node Identification: Nodes in the Bayesian Network represent key factors influencing educational outcomes. These include variables such as:

- **Student Attributes:** Hours_Studied, Motivation_Level, Attendance, Previous_Scores.
- **Environmental Factors:** Parental_Involvement, Access_to_Resources, Family_Income.
- **Institutional Factors:** Teacher_Quality, School_Type, Tutoring_Sessions.
- **Behavioral Factors:** Physical_Activity, Sleep_Hours, Extracurricular_Activities.

The target variable was identified as Exam_Score, with three possible outcomes: Fail, Pass, and Excellent.

2) Edge Specification: Edges between nodes were established based on expert knowledge and real-world instances. For example:

- Hours_Studied and Motivation_Level influence Exam_Score.
- Teacher_Quality impacts Hours_Studied and Motivation_Level.
- Family_Income affects Access_to_Resources and Parental_Involvement.

These connections reflect causal relationships or dependencies that align with domain expertise and practical observations.

B. Learning Parameters

Learning parameters was majorly done through data. Our dataset was taken from Kaggle and contained over 6000 records.

1) Data Preprocessing: To ensure that the Bayesian Network could operate effectively on the dataset, several preprocessing steps were performed:

1) **Handling Missing Values:** Rows with missing values were dropped to preserve data integrity.
2) **Discretization:** Many variables, such as Hours_Studied and Previous_Scores, were initially continuous. These were discretized into categories (Low, Medium, High) based on logical thresholds or domain expertise. This step simplifies the construction of Conditional Probability Tables (CPTs) and reduces computational complexity.

2) Conditional Probability Tables (CPTs):

- For nodes with no parents, prior probabilities were directly estimated from the data.
- For nodes with parents, the CPTs were learned using the Maximum Likelihood Estimator (MLE). MLE computes probabilities by maximizing the likelihood of the observed data given the Bayesian Network structure.

To address the issue of zero probabilities in CPTs (arising from unseen combinations of parent and child states in the training data), **Laplacian smoothing** was applied. For each CPT, the number of possible states of the variable (denoted as k) was determined, and a smoothing factor of α/k was added uniformly to every entry.

While expert knowledge defined the initial structure, MLE refined the relationships by adjusting CPT values based on the training dataset, ensuring the model accurately reflects empirical patterns.

C. Inference

Inference in a Bayesian Network involves querying the model to estimate the probabilities of target variables given evidence (known values of other variables). In this study, the focus was on predicting Exam_Score based on observed features.

During inference, evidence variables such as Hours_Studied, Motivation_Level, and Attendance were observed for each test case. For example:

- **Query:** What is the most likely Exam_Score given that Hours_Studied = Low and Teacher_Quality = High?

The Bayesian Network calculates the posterior probabilities for Fail, Pass, and Excellent.

D. Evaluation Metrics

The model was then evaluated based on the following metrics:

- **Accuracy:** Percentage of correctly classified test cases.
- **Precision:** Proportion of true positive predictions among all positive predictions.
- **Recall:** Proportion of actual positive cases correctly identified by the model.
- **F1-Score:** Harmonic mean of precision and recall, balancing their trade-offs.

These metrics were computed for each class (Fail, Pass, Excellent) and overall performance.

E. Summary

Overall, the network handled a relatively large number of nodes and dependencies, demonstrating its scalability for complex systems.

IV. Results

The model achieved the following metrics on the test dataset:

- **Accuracy:** 84.74%

A breakdown of performance across the three target classes is shown in Table I.

Class	Precision	Recall	F1-Score
Excellent	0.08	0.08	0.08
Fail	0.83	0.65	0.73
Pass	0.90	0.92	0.91

TABLE I
PERFORMANCE METRICS ACROSS THE TARGET CLASSES.

V. Discussion and Analysis

The results demonstrate the effectiveness of integrating domain expertise with data-driven learning in Bayesian Networks. The expert-defined Bayesian Network achieved an accuracy of 84.79%, demonstrating a balance between interpretability and predictive performance. However, the inability of the model to predict the 'Excellent' class effectively indicates a significant limitation likely stemming from class imbalance or insufficient representation of critical features for this category.

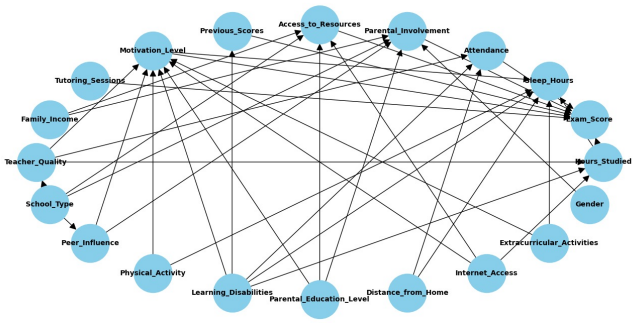

Fig. 1. Bayesian Network Model Structure - Expert Defined.

This was the structure learned through domain knowledge. It integrates expert knowledge with real-world observations, ensuring the network is both interpretable and grounded in reality. The directed edges form a Directed Acyclic Graph (DAG), which enables efficient probabilistic inference by leveraging the conditional independencies encoded in the structure. Given such a complex network, we worked on getting probabilities through data. We started by splitting the dataset into a ratio of 60/30, using 60% of the samples for generating the Conditional Probability Tables (CPTs) and the remaining 30% for testing the Bayesian Network. While testing, each row from the test data was provided as evidence

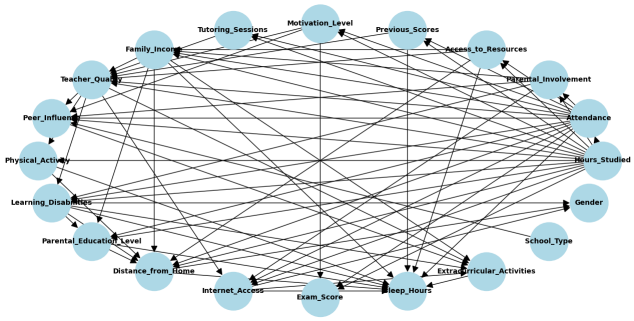

Fig. 2. Bayesian Network Model Structure - K2.

to predict an *Exam Score*. The final predictions were then evaluated for their accuracy using various metrics.

This was the structure learned through the structural learning algorithm.

A. Comparison of Expert-Defined and K2 Structure Learning

Bayesian Networks allow for the flexibility of structure definition through both expert input and data-driven techniques like K2 structure learning. We explored both approaches, and their results provide valuable insights:

1) K2 Structure Learning: The K2 structure learning method involves automatically learning the structure of the Bayesian Network from data based on a predefined node order. Using K2, the model achieved a higher accuracy of 88.53%. However, the structure produced by this method was highly complex, with many additional edges connecting nodes. This complexity made the model difficult to interpret, especially for stakeholders or domain experts.

The metrics for the K2 model were:

- **Accuracy:** 88.53%.
- **Precision (Fail):** 88%.
- **Recall (Fail):** 56%.
- **Precision (Pass):** 89%.
- **Recall (Pass):** 99%.

While the model performed well for the ”Pass” class, it completely failed to predict the ”Excellent” class (Precision: 0%, Recall: 0%). This highlights a critical weakness: K2 structure learning often prioritizes accuracy for majority classes at the expense of minority classes.

2) Expert-Defined Structure: The expert-defined structure was simpler and more interpretable. It relied on domain knowledge to define key dependencies, ensuring that causal relationships aligned with real-world observations. This model achieved an accuracy of 84.79% with a balanced performance across all classes.

The metrics for the expert-defined model were:

- **Accuracy:** 84.79%.
- **Precision (Fail):** 83%.
- **Recall (Fail):** 65%.
- **Precision (Pass):** 90%.
- **Recall (Pass):** 92%.
- **Precision (Excellent):** 8%.
- **Recall (Excellent):** 8%.

Although slightly less accurate than the K2 model, the expert-defined model demonstrated better performance in identifying minority classes, especially the ”Fail” and ”Excellent” categories. This balance is crucial in educational settings where understanding failing students and outliers is as important as identifying the majority.

3) Rationale for Choosing the Expert-Defined Model: After a thorough comparison, we chose the expert-defined structure for the following reasons:

- **Interpretability:** The simpler structure aligns with domain knowledge, making it more comprehensible to stakeholders.
- **Balanced Performance:** The expert-defined model showed better recall for minority classes, making it more reliable for nuanced analysis.
- **Robustness:** By leveraging domain expertise, the model avoided overfitting, a common issue with data-driven approaches like K2.

B. Advantages of Bayesian Networks

This study highlights several strengths of Bayesian Networks:

- **Flexibility:** Bayesian Networks can incorporate both expert knowledge and data-driven learning, enabling a tailored approach to problem-solving.
- **Causal Interpretability:** The directed edges in Bayesian Networks explicitly represent causal relationships, making them particularly valuable in educational settings. For instance, educators can observe how variations in ’Motivation Level’ or ’Hours Studied’ directly influence ’Exam Score,’ enabling data-informed decision-making. The ability to update probabilities dynamically based on new evidence ensures that the model remains relevant for real-time monitoring and prediction.
- **Scalability:** The model successfully handled over 6000 records, demonstrating its suitability for large datasets.
- **Iterative Learning:** The structure and parameters can be refined iteratively with new data or expert input, ensuring the model remains relevant over time.

By integrating domain expertise and data-driven learning, Bayesian Networks offer a powerful framework for modeling complex systems with interpretable and actionable insights.

VI. Conclusion and Future Work

Compared to traditional machine learning approaches (e.g., Decision Trees, Neural Networks), the Bayesian Network offers interpretability by explicitly modeling causal relationships. The network structure also allows for iterative updates based on new evidence or expert feedback. One drawback of the model is that the model's performance is heavily dependent on data quality and the number of samples. Despite the size of the dataset (over 6,000 records), the distribution of target classes was imbalanced, with the 'Excellent' category significantly underrepresented. This imbalance skewed the model's ability to predict minority classes effectively, as evidenced by the low precision and recall for 'Excellent.' Future work could address this limitation using techniques such class weights and domain knowledge [13], to ensure balanced learning.

References

[1] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge, UK: Cambridge University Press, 2000.B. Rieder, Engines of Order: A Mechanology of Algorithmic Techniques. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.

[2] D. Karaboğa and E. Demir, "Bayesian Network Modeling of Academic Success Factors Using PISA Data," Educational Research and Evaluation, 2023.

[3] Pintrich, P. R., Schunk, D. H. Motivation in Education: Theory, Research, and Applications, 2002.

[4] Ryan, R. M., Deci, E. L. "Self-Determination Theory and the Facilitation of Intrinsic Motivation," American Psychologist, 55(1), 68-78, 2000.

[5] Kolb, D. A. Experiential Learning: Experience as the Source of Learning and Development, 2014.

[6] Dunn, R., Dunn, K. Teaching Secondary Students through Their Individual Learning Styles, 1993.

[7] Coleman, J. S., et al. Equality of Educational Opportunity, 1966.

[8] Fan, X., Chen, M. "Parental Involvement and Students' Academic Achievement," Educational Psychology Review, 13(1), 1-22, 2001.

[9] Reardon, S. F. "The Widening Academic Achievement Gap," Community Investments, 23(2), 19-39, 2011.

[10] Duncan, G. J., Magnuson, K. "The Nature and Impact of Early Achievement Skills," Whither Opportunity, 47-70, 2011.

[11] Lubienski, C., Lubienski, S. T. The Public School Advantage, 2014.

[12] I. Nunes and P. Lemaire, "Bayesian Networks in Educational Research: Methodological Advances and Practical Applications," Educational Technology Research and Development, vol. 67, no. 2, pp. 189-212, 2019.

[13] R. Pansanga, "Improving precision and recall in machine learning: Tips and techniques," Medium, 2023